

# PROGRAM FOR ALLOWING COMPUTER TO CARRY OUT AUTOMATIC SELECTION OF TRANSLATION SYSTEM, AND COMPUTER-READABLE RECORDING MEDIUM RECORDING THE PROGRAM

Patent number: JP2003263434  
Publication date: 2003-09-19  
Inventor: YASUDA YOSHIYUKI; SUGAYA FUMIAKI; TAKEZAWA TOSHIYUKI; YAMAMOTO SEIICHI  
Applicant: ATR ADVANCED TELECOMM RES INST  
Classification:  
- international: G06F17/28; G06F17/28; (IPC1-7): G06F17/28  
- european:  
Application number: JP20020065365 20020311  
Priority number(s): JP20020065365 20020311

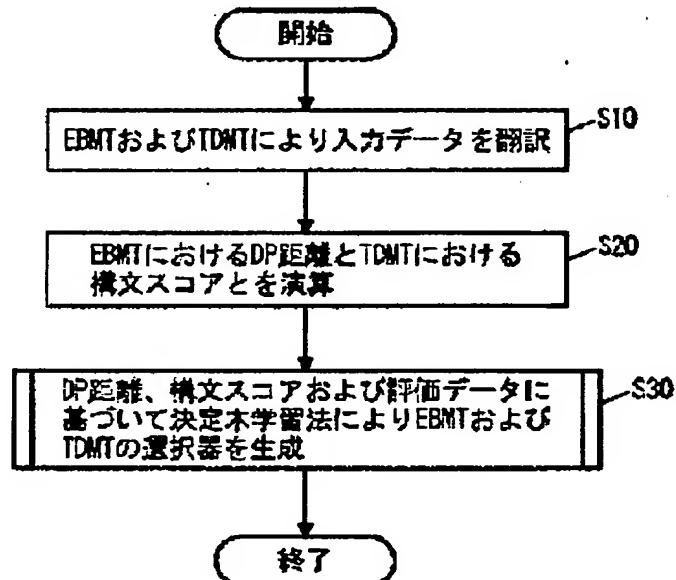
[Report a data error here](#)

## Abstract of JP2003263434

**PROBLEM TO BE SOLVED:** To provide a program for allowing a computer to carry out automatic selection of a translation system suitable for the translation of input data, out of a plurality of translation systems.

**SOLUTION:** Input data is translated by two translation systems TDMT and EBMT (step S10). A sentence structure score showing similarity between the input data and examples in translating the input data by the EBMT, and a DP distance showing similarity between the input data and examples in translating the input data by the EBMT, are computed (step S20). Evaluation data showing whether the TDMT and EBMT are suitable for the translation of the input data, and the sentence structure score and DP distance computed in the step S20, are used to generate a selector for selecting the translation system suitable for the translation of the input data by a decision tree learning method (step S30).

**COPYRIGHT:** (C)2003,JPO



Data supplied from the [esp@cenet](mailto:esp@cenet) database - Worldwide

BEST AVAILABLE COPY

(19)日本国特許庁 (JP)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2003-263434

(P2003-263434A)

(43)公開日 平成15年9月19日 (2003.9.19)

(51)Int.Cl.

G 0 6 F 17/28

識別記号

F I

テーマコード(参考)

G 0 6 F 17/28

Z 5 B 0 9 1

審査請求 未請求 請求項の数 6 O.L. (全 12 頁)

(21)出願番号 特願2002-65365(P2002-65365)

(22)出願日 平成14年3月11日 (2002.3.11)

(71)出願人 393031586

株式会社国際電気通信基礎技術研究所  
京都府相楽郡精華町光台二丁目2番地2

(72)発明者 安田 圭志

京都府相楽郡精華町光台二丁目2番地2  
株式会社国際電気通信基礎技術研究所内

(72)発明者 皆谷 史昭

京都府相楽郡精華町光台二丁目2番地2  
株式会社国際電気通信基礎技術研究所内

(74)代理人 100064746

弁理士 深見 久郎 (外4名)

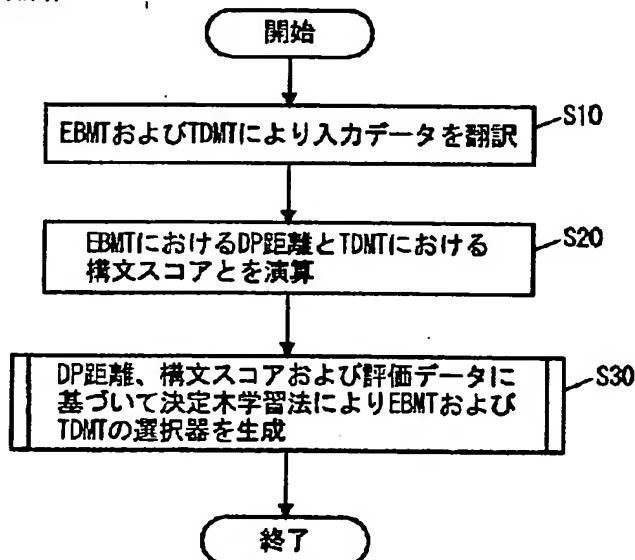
最終頁に続く

(54)【発明の名称】 翻訳システムの自動選択をコンピュータに実行させるためのプログラム、およびそのプログラムを記録したコンピュータ読み取り可能な記録媒体

(57)【要約】

【課題】 複数の翻訳システムから入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムを提供する。

【解決手段】 2つの翻訳システム TDMT, EBMT によって入力データを翻訳する (ステップ S 10)。そして、入力データを TDMT により翻訳する際の入力データと用例との類似性を示す構文スコアと、入力データを EBMT により翻訳する際の入力データと用例との類似性を示す DP 距離とを演算する (ステップ S 20)。入力データの翻訳に TDMT, EBMT が適しているか否かを示す評価データと、ステップ S 20において演算した構文スコアおよび DP 距離とを用いて決定木学習法により入力データの翻訳に適した翻訳システムを選択するための選択器を生成する (ステップ S 30)。



## 【特許請求の範囲】

【請求項1】  $m$  ( $m$ は自然数) 個の翻訳システムから  $n$  ( $n$ は自然数) 個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムであって、

前記  $m$  個の翻訳システムの各々により前記  $n$  個の入力データの各々を翻訳する第1のステップと、

前記  $m$  個の翻訳システムの各々が前記  $n$  個の入力データの各々を翻訳する際の翻訳し易さを示す  $n$  個の指標を前記  $m$  個の翻訳システムに対して演算する第2のステップと、

前記演算された第1番目の  $n$  個の指標乃至第  $m$  番目の  $n$  個の指標と、前記  $n$  個の入力データの翻訳に適した翻訳システムを前記  $n$  個の入力データの各入力データごとに示す  $n$  個の評価データとに基づいて、決定木学習法により前記  $m$  個の翻訳システムの各々に対して翻訳の適合性を演算する第3のステップとをコンピュータに実行させるためのプログラム。

【請求項2】 前記  $m$  個の翻訳システムは、第1および第2の翻訳システムであり、

前記第1のステップは、

前記第1の翻訳システムを用いて、前記  $n$  個の入力データの各々を構成する各単語を翻訳語に変換する第1のサブステップと、

前記第2の翻訳システムを用いて、前記  $n$  個の入力データを構文構造によって分割した複数のパターンの各々を翻訳語に変換する第2のサブステップとを含み、

前記第2のステップは、

前記各単語を前記翻訳語に変換する際の前記各単語の意味と各単語に対応する翻訳語の意味との類似性を示すDP距離を前記  $n$  個の入力データの各々に対して演算する第3のサブステップと、

前記パターンを前記翻訳語に変換する際の前記パターンの意味と前記翻訳語の意味との類似性を示す構文スコアを前記  $n$  個の入力データの各々に対して演算する第4のサブステップとを含む、請求項1に記載のコンピュータに実行させるためのプログラム。

【請求項3】 前記第1のステップにおいて前記第1および第2の翻訳システムにより翻訳されたそれぞれ第1および第2の翻訳データを外部へ出力する第4のステップをさらにコンピュータに実行させる、請求項2に記載のコンピュータに実行させるためのプログラム。

【請求項4】 前記  $n$  個の評価データを外部から受付ける第5のステップをさらにコンピュータに実行させる、請求項3に記載のコンピュータに実行させるためのプログラム。

【請求項5】 前記評価データは、前記第1および第2の翻訳データに基づいて前記翻訳の正しさを人間により評価したデータである、請求項4に記載のコンピュータに実行させるためのプログラム。

【請求項6】 請求項1から請求項5のいずれか1項に記載のプログラムを記録したコンピュータ読取り可能な記録媒体。

## 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】 この発明は、複数の翻訳システムから翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラム、およびそのプログラムを記録したコンピュータ読取り可能な記録媒体に関するものである。

## 【0002】

【従来の技術】 今までに、様々な機械翻訳システムが研究・開発されている。例えば、EBMT (Example Based Machine Translation)、およびTDMT (Transfer Driven Machine Translation) が研究・開発されている。

【0003】 EBMTは、1つの文に含まれる単語単位で変換すべき翻訳語を検索して入力された入力データを翻訳する翻訳システムである。また、TDMTは、1つの文よりも短い、構文構造の基本単位である構成素境界パターンを単位として翻訳語に変換する翻訳システムである。

## 【0004】

【発明が解決しようとする課題】 しかし、これらの翻訳システムは、それぞれ、特定のドメインまたは特定の表現形式に適しているため、ある翻訳システムにより入力データを翻訳する翻訳器に、その翻訳システムによる翻訳に適さない入力データが入力されると、正確な翻訳ができないという問題が生じる。

【0005】 そこで、この発明は、かかる問題を解決するためになされたものであり、その目的は、複数の翻訳システムから入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムを提供することである。

【0006】 また、この発明の別の目的は、複数の翻訳システムから入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムを記録したコンピュータ読取り可能な記録媒体を提供することである。

## 【0007】

【課題を解決するための手段および発明の効果】 この発明によれば、 $m$  ( $m$ は自然数) 個の翻訳システムから  $n$  ( $n$ は自然数) 個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムは、 $m$  個の翻訳システムの各々により  $n$  個の入力データの各々を翻訳する第1のステップと、 $m$  個の翻訳システムの各々が  $n$  個の入力データの各々を翻訳する際の翻訳し易さを示す  $n$  個の指標を  $m$  個の翻訳システムに対して演算する第2のステップと、演算された第1番目

の  $n$  個の指標乃至第  $m$  番目の  $n$  個の指標と、 $n$  個の入力データの翻訳に適した翻訳システムを  $n$  個の入力データの各入力データごとに示す  $n$  個の評価データとに基づいて、決定木学習法により  $m$  個の翻訳システムの各々に対して翻訳の適合性を演算する第 3 のステップとをコンピュータに実行させるためのプログラムである。

【0008】好ましくは、 $m$  個の翻訳システムは、第 1 および第 2 の翻訳システムであり、第 1 のステップは、第 1 の翻訳システムを用いて、 $n$  個の入力データの各々を構成する各単語を翻訳語に変換する第 1 のサブステップと、第 2 の翻訳システムを用いて、 $n$  個の入力データを構文構造によって分割した複数のパターンの各々を翻訳語に変換する第 2 のサブステップとを含み、第 2 のステップは、各単語を翻訳語に変換する際の各単語の意味と各単語に対応する翻訳語の意味との類似性を示す DP 距離を  $n$  個の入力データの各々に対して演算する第 3 のサブステップと、パターンを翻訳語に変換する際のパターンの意味と翻訳語の意味との類似性を示す構文スコアを  $n$  個の入力データの各々に対して演算する第 4 のサブステップとを含む。

【0009】より好ましくは、 $m$  ( $m$  は自然数) 個の翻訳システムから  $n$  ( $n$  は自然数) 個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムは、第 1 のステップにおいて第 1 および第 2 の翻訳システムにより翻訳されたそれぞれ第 1 および第 2 の翻訳データを外部へ出力する第 4 のステップをさらにコンピュータに実行させる。

【0010】さらに好ましくは、 $m$  ( $m$  は自然数) 個の翻訳システムから  $n$  ( $n$  は自然数) 個の入力データの翻訳に適した翻訳システムの自動選択をコンピュータに実行させるためのプログラムは、 $n$  個の評価データを外部から受付ける第 5 のステップをさらにコンピュータに実行させる。

【0011】さらに好ましくは、評価データは、第 1 および第 2 の翻訳データに基づいて翻訳の正しさを人間により評価したデータである。

【0012】また、この発明によれば、プログラムを記録したコンピュータ読取り可能な記録媒体は、請求項 1 から請求項 5 のいずれか 1 項に記載のプログラムを記録したコンピュータ読取り可能な記録媒体である。

【0013】従って、この発明によれば、複数の翻訳システムから入力データの翻訳に適した翻訳システムを選択することができる。

【0014】

【発明の実施の形態】本発明の実施の形態について図面を参照しながら詳細に説明する。なお、図中同一または相当部分には同一符号を付してその説明は繰返さない。

【0015】この発明によるプログラムは、複数の翻訳システムから入力データの翻訳に適した翻訳システムを自動選択する。そして、この発明の実施の形態において

は、2 つの翻訳システムから入力データの翻訳に適した翻訳システムを自動選択するプログラムについて説明し、2 つの翻訳システムは、TDMT および EBMT である。

【0016】図 1 は、2 つの翻訳システム (TDMT および EBMT) の概要を説明するための図である。EBMT では、多くの用例を含むコーパスから、各入力データ ( $D_1, D_2, \dots, D_n$ ) に最も類似した用例を検索し、最も類似した用例を翻訳に用いる。コーパスは、多くの原言語用例 (入力側の言語) と目的言語用例 (出力側言語) のペアから成る。最も類似した用例は、各入力とコーパス内の各原言語用例との DP 距離に基づいて検索される。そして、DP 距離は次式で定義される。

【0017】

$$P_{DP} = \frac{I + D + 2 \sum D_{semantic}}{L_{input} + L_{example}}$$

【数 1】

①  $P_{DP}$   $\leftarrow$  最も少なく意味距離の和が最小のコーパスを選択する。

【0018】但し、 $L_{input}$  は入力を表わし、 $L_{example}$  は、コーパス内の用例の単語数を表わし、 $I$  は、入力文とコーパス内の用例とを DP マッチングで比較した時の挿入語数を表わし、 $D$  は、入力文とコーパス内の用例とを DP マッチングで比較した時の脱落語数を表わし、 $D_{semantic}$  は、単語間の意味距離である。

【0019】式 (1) により定義される DP 距離  $P_{DP}$  は、入力文と用例とがどの程度類似しているかを示し、「0」から「1」までの値をとる。そして、入力文と用例とが完全に一致した場合、DP 距離  $P_{DP}$  は「0」となる。

【0020】従って、入力文  $D_1 \sim D_n$  の各々について、最も類似した用例の各々をコーパスから抽出するとき、コーパスに含まれる用例の全てについて式 (1) を用いて DP 距離  $P_{DP}$  を演算し、DP 距離  $P_{DP}$  が最も小さい用例をコーパスから抽出する。

【0021】翻訳の手順を以下に述べる。ここで、ある入力データ  $D_i$  ( $1 \leq i \leq n$ ) とし、 $D_i$  は単語  $W_1, \dots, W_p$  ( $p$  は自然数) の系列から成るとする。コーパスの検索により得られた原言語用例を  $G$  とし、 $G$  は単語  $WG_1, \dots, WG_p$  の系列からなるとする。この原言語文に対応する目的言語用例を  $M$  とし、 $M$  は単語  $WM_1, \dots, WM_p$  の系列から成るとする。入力データ  $D_i$  と、原言語用例  $G$  とが完全に一致した場合は、原言語用例に対応する目的言語用例  $M$  をそのまま出力する。入力データ  $D_i$  と、原言語用例  $G$  とが完全に一致しない場合は、まず、入力データ  $D_i$  と原言語用例  $G$  を比較して異なる単語を検出する。ここで、入力データ  $D_i$  の単語  $W_j$  ( $1 \leq j \leq p$ ) と目的言語用例  $G$  の単語  $WG_j$  のみが異なり、それ以外の単語については同じであるとすると、 $WG_j$  に対応する単語を辞書などの情報を用いて  $W_j$  に対応する目的言語の単語  $WM_{new}$  を

【0022】

【0023】

求め、目的言語用例Mにおいて単語WM<sub>k</sub>を単語WM<sub>new<sub>k</sub></sub>に置き換えて出力する。

【0022】TDMTは、入力データDの各データD<sub>1</sub>, D<sub>2</sub>, …, D<sub>n</sub>よりも短く、かつ、構文構造の基本単位である構成素境界パターン（以下、「パターン」と言う。）を単位としてコーパスから変換知識を学習し、その学習した変換知識を用いて入力データDを翻訳する翻訳システムである。すなわち、入力データD (D<sub>1</sub>, D<sub>2</sub>, …, D<sub>n</sub>) は、q (qは自然数) 個のパターンPT<sub>1</sub>～PT<sub>q</sub>に分割され、パターンPT<sub>1</sub>～PT<sub>q</sub>の各々を、コーパスから選択したパターンTP<sub>1</sub>～TP<sub>q</sub>にそれぞれ変換することにより、入力データDは翻訳される。

【0023】より具体的には、パターンPT<sub>1</sub>～PT<sub>q</sub>の各々をパターンTP<sub>1</sub>～TP<sub>q</sub>に変換するとき、次式により定義される構文スコアを演算する。

【0024】

【数2】

$$P_{TDMT} = \underset{(Pj) \in P}{\operatorname{argmin}} \sum S \cdot (b_i, Pj) \quad \cdots (2)$$

【0025】但し、PはTDMTが持つパターンの集合を表わし、b<sub>i</sub> (iは自然数) は入力文を解析して得られる部分文を表わし、Sはパターンと部分文との意味距離を表わす。

【0026】構文スコアP<sub>TDMT</sub>は、「0」以上の値を探り、値が小さい方が意味の近いパターンであることを示す。

【0027】従って、パターンPT<sub>1</sub>～PT<sub>q</sub>の各々について、コーパスに含まれる全ての用例について式(2)を用いて構文スコアP<sub>TDMT</sub>を演算し、構文スコアP<sub>TDMT</sub>が最小となるパターンがコーパスから選択される。

【0028】このように、TDMTは、入力データDの各データD<sub>1</sub>, D<sub>2</sub>, …, D<sub>n</sub>よりも短いパターンを基本単位として入力データDを翻訳する翻訳システムである。

【0029】なお、TDMTでは、複数の文からなる長い発話や、話し言葉に見られる文法法規から逸脱した入力等の原因で、変換を適用する部分文の組合せにより1つの構文木を作れなかった場合は、最も整合性のとれた部分に分割し、部分毎に翻訳を行なう。

【0030】図2は、この発明によるプログラムが、上述した2つの翻訳システムから入力データの翻訳に適した翻訳システムを自動選択する際の概念を示す概念図である。

【0031】まず、学習文10がTDMT翻訳器20およびEBMT翻訳器30に入力される。TDMT翻訳器20は、学習文10を上述したTDMTにより翻訳し、翻訳文であるTDMT出力と、構文スコアP<sub>TDMT</sub>とを出力する。また、EBMT翻訳器30は、学習文10を上

述したEBMTにより翻訳し、翻訳文であるEBMT出力と、DP距離P<sub>DP</sub>とを出力する。

【0032】評価器40は、TDMT翻訳器20からのTDMT出力と、EBMT翻訳器30からのEBMT出力とを受け、TDMT出力およびEBMT出力を、翻訳一対比較法により評価する。つまり、人間が学習文10の翻訳文としてTDMT出力とEBMT出力とのうち、どちらが適しているかを翻訳前の学習文10と翻訳文であるTDMT出力（またはEBMT出力）とを一対一に比較して評価する。そして、評価器40は、評価結果を出力する。

【0033】決定木学習器50は、TDMT翻訳器20から構文スコアP<sub>TDMT</sub>を受け、EBMT翻訳器30からDP距離P<sub>DP</sub>を受け、評価器40から評価結果を受けれる。そして、決定木学習器50は、構文スコアP<sub>TDMT</sub>、DP距離P<sub>DP</sub>、および評価結果に基づいて、決定木学習法を用いて選択器を生成する。

【0034】自動選択器60は、決定木学習器50から選択器に基づいて、学習文10の翻訳に適した翻訳システムを選択する。

【0035】このように、この発明によるプログラムは、TDMTにおいて変換すべきパターンの見つけ易さ、すなわち、翻訳し易さを示す構文スコアP<sub>TDMT</sub>、EBMTにおいて変換すべき単語の見つけ易さ、すなわち、翻訳のし易さを示すDP距離P<sub>DP</sub>、および評価結果に基づいて、適した翻訳システムを選択するための選択器を生成することを特徴とする。

【0036】図3は、この発明によるプログラムがコンピュータに実行させるためのステップを説明するためのフローチャートである。動作が開始されると、EBMTおよびTDMTにより入力データを翻訳する（ステップS10）。そして、EBMTにおけるDP距離P<sub>DP</sub>と、TDMTにおける構文スコアP<sub>TDMT</sub>とを演算し（ステップS20）、構文スコアP<sub>TDMT</sub>、DP距離P<sub>DP</sub>、および評価結果に基づいて、決定木学習法を用いて選択器を生成する（ステップS30）。これにより、一連の動作は終了する。

【0037】図4は、図3に示すフローチャートのステップS30における決定木学習法による選択器の生成の詳細な動作を説明するためのフローチャートである。図4に示すフローチャートについて説明する前に、前提について説明する。学習データLは、TDMTにおける構文スコアP<sub>TDMT</sub>であるTと、EBMTにおけるDP距離P<sub>DP</sub>であるEと、評価データであるCとから成る。すなわち、L = (T, E, C) である。

【0038】また、構文スコアT、DP距離E、および評価データCは、それぞれ、n個の要素から成る。すなわち、T = |t<sub>1</sub>, t<sub>2</sub>, …, t<sub>n</sub>|、E = |e<sub>1</sub>, e<sub>2</sub>, …, e<sub>n</sub>|、C = |c<sub>1</sub>, c<sub>2</sub>, …, c<sub>n</sub>| である。そして、T = |t<sub>1</sub>, t<sub>2</sub>, …, t<sub>n</sub>|、E =

$|e_1, e_2, \dots, e_n|$ 、および  $C = |c_1, c_2, \dots, c_n|$  の各要素は相互に対応関係を有する。例えば、 $t_1, e_1$ 、および  $c_1$  は相互に対応関係を有する。その他の要素についても同様である。

【0039】さらに、評価データ  $C$  の要素  $c_1, c_2, \dots, c_n$  は、「0」か「1」のいずれかの値を探る。そして、「0」は TD M T の翻訳システムが優勢であることを示し、「1」は EB M T の翻訳システムが優勢であることを示す。

【0040】なお、実際には、2つの翻訳システム TD M T, EB M T を比較した場合、同等と評価される場合もあるが、このようなデータは学習データ L に用いられない。

【0041】図3に示すステップ S 20 の後、学習データ  $L = (T, E, C)$  が渡される（ステップ S 31）。そして、構文スコア  $T$  の全てのメンバー  $t_1, t_2, \dots, t_n$  の値を閾値の候補として、構文スコア  $T$  に対する利得比を求める。この利得比の求め方について後述する。

【0042】利得比の演算が終了すると、全てのメンバー  $t_1, t_2, \dots, t_n$  を閾値とした場合の構文スコア  $T$  に対する利得比の中から最大となる利得比  $T_{gain}$  を抽出し、 $T_{gain}$  を与える閾値  $T_{ht}$  を抽出する。

【0043】DP 距離  $E$  についても、構文スコア  $T$  の場合と同様に最大となる利得比  $E_{gain}$  と、 $E_{gain}$  を与える閾値  $T_{he}$  を求める。そして、 $T_{gain}$  を  $E_{gain}$  と比較し、大きい方を最大利得比  $MAX_{gain}$  とする（ステップ S 32）。

【0044】ここで、利得比の演算について説明する。例として  $T$  に対する利得比を、メンバー  $t_k$  を閾値として求める場合について説明する。 $T$  の値により、 $T$  と、それに対応する  $C$  をソートし、 $t_k$  を境界として  $C = |c_1, c_2, \dots, c_n|$  を2つに分割する。分割後の  $C$  をそれぞれ  $C_{div(1)}$  と  $C_{div(2)}$  とで表わす（ $C = C_{div(1)} + C_{div(2)}$ ）。

【0045】ここで、値が「0」となる  $C$  のメンバーの数を  $freq(C_0, C)$  で表わし、値が「1」となる  $C$  のメンバーの数を  $freq(C_1, C)$  で表わす。また、 $|C|$  を  $C$  に含まれるメンバーの数とすると、 $C$  の平均情報量は次式により求められる。

【0046】

【数3】

$$info(c) = -\sum_{j=0}^1 \frac{freq(C_j, C)}{|C|} \times \log_2 \left( \frac{freq(C_j, C)}{|C|} \right) \quad \dots (3)$$

【0047】info(C) は、C 内にある 1 つの事例に属するクラスを道程するのに必要な情報量の平均値となる。

【0048】また、分割後のデータ  $C_{div(1)}$ 、 $C_{div(2)}$  に同様な評価を考えると、部分集合上で荷重平均をとって、次式により求められる。

【0049】

【数4】

$$info_{tk}(c) = \sum_{j=0}^1 \frac{|C_{div(j)}|}{|C|} \times info(C_{div(j)}) \quad \dots (4)$$

【0050】さらに、分割情報量は次式により求められる。

【0051】

【数5】

$$split\_info(t_k) = -\sum_{j=0}^1 \frac{|C_{div(j)}|}{|C|} \times \log_2 \frac{|C_{div(j)}|}{|C|} \quad \dots (5)$$

【0052】そして、利得比基準は、これまでの式を用いて次式により求められる。

【0053】

【数6】

$$gain\_ratio(t_k) = \frac{info(c) - info_{tk}(c)}{split\_info(t_k)} \quad \dots (6)$$

【0054】従って、ステップ S 32においては、上述した式 (3) ~ (6) を用いて構文スコア  $T$  および DP 距離  $E$  に対して利得比が演算され、最終的に、最大利得比  $MAX_{gain}$  が演算される。

【0055】ステップ S 32 の後、最大利得比  $MAX_{gain}$  が「0」以下、または  $C$  のメンバー  $c_1, c_2, \dots, c_n$  の値が全て同じか否かを判定する（ステップ S 33）。そして、最大利得比  $MAX_{gain}$  が「0」以下である場合、または  $C$  のメンバー  $c_1, c_2, \dots, c_n$  の値が全て同じである場合、決定木学習法により選択器を生成する動作は終了する（ステップ S 34）。

【0056】ステップ S 33において、最大利得比  $MAX_{gain}$  が「0」よりも大きく、かつ、 $C$  のメンバー  $c_1, c_2, \dots, c_n$  の値が相互に不一致である場合、学習データ L は 2 つの学習データ  $L_0, L_1$  に分割される（ステップ S 35）。この場合、ステップ S 32における最大利得比  $MAX_{gain} = T_{gain}$  であるなら、閾値  $T_{ht}$  を用いて学習データ L を学習データ  $L_0, L_1$  に分割し、ステップ S 32における最大利得比  $MAX_{gain} = E_{gain}$  であるなら、閾値  $T_{he}$  を用いて学習データ L を学習データ  $L_0, L_1$  に分割する。

【0057】この分割の具体例を図5に示す。閾値  $T_{ht} = t_k$  であるとき  $T = |t_1, t_2, \dots, t_k, \dots, t_n|$  を閾値  $t_k$  により 2 つのデータ  $T_0 = |t_1, t_2, \dots, t_k|$ 、 $T_1 = |t_{k+1}, t_2, \dots, t_n|$  に分割する。また、 $E = |e_1, e_2, \dots, e_k, \dots, e_n|$  を閾値  $t_k$  に対応するメンバー  $e_k$  を基準にして 2 つのデータ  $E_0 = |e_1, e_2, \dots, e_k|$ 、 $E_1 = |e_{k+1}, e_2, \dots, e_n|$  に分割する。さらに、 $C = |c_1, c_2, \dots, c_k, \dots, c_n|$  についても、同様に、2 つのデータ  $C_0 = |c_1, c_2, \dots, c_k|$ 、 $C_1 = |c_{k+1}, c_2, \dots, c_n|$  に分割する。

【0058】従って、学習データ  $L_0 = (|t_1, t_2, \dots, t_k|, |e_1, e_2, \dots, e_k|, |c_1, c_2, \dots, c_k|)$

…  $c_k$  ) であり、学習データ  $L_1 = ( |t_{k+1}, t_2, \dots, t_n|, |e_{k+1}, e_2, \dots, e_n|, |c_{k+1}, c_2, \dots, c_n| )$  である。

【0059】閾値  $T_{HE}$  を用いて学習データ  $L$  を分割する場合も同様である。図4に示すフローチャートのステップS35において、学習データ  $L$  が学習データ  $L_0, L_1$  に分割されると、学習データ  $L_0, L_1$  の各々についてステップS31～S35が繰返し行なわれる（ステップS40, S50）。

【0060】ステップS40, S50の各々に含まれるステップS33において、最大利得比  $MAX_{gain}$  が「0」よりも大きく、かつ、Cのメンバー  $c_1, c_2, \dots, c_n$  の値が相互に不一致である場合、学習データ  $L_0, L_1$  の各々は、2つの学習データに分割される。そして、分割された各学習データを対象としてステップS31～S35が繰返し行なわれる。

【0061】すなわち、学習データ  $L$  を最初の学習データとしていずれの翻訳システムが適しているかの判定を開始し、ツリー状に学習データ  $L$  を分割し、その分割した学習データの各々を対象としていずれの翻訳システムが適しているかの判定を行なう。

【0062】そして、各学習データを対象とした判定のステップS33において、最大利得比  $MAX_{gain}$  が「0」以下である場合、またはCのメンバー  $c_1, c_2, \dots, c_n$  の値が全て同じである場合に決定木学習法により選択器を生成する動作は終了する。

【0063】選択器を生成することは、構文スコア（「TDMT距離」とも言う。）とEBM距離（「EBMT距離」とも言う。）によって形成される二次元空間において、TDMTに適した領域および／またはEBMTに適した領域に分割することに相当する。

【0064】例えば、図6に示すように、点A-B-C-Dによって囲まれる領域を、点E-F-G-H-I-J-D-Cによって囲まれる領域と、点A-B-E-F-G-H-I-Jによって囲まれる領域とに分割する。

【0065】図4に示すフローチャートのステップS33において、Cのメンバー  $c_1, c_2, \dots, c_n$  の値が全て同じであるとき、TDMTおよびEBMTのいずれか一方が入力データの翻訳に適した翻訳システムであることを示すので、点A-B-C-Dによって囲まれる領域は分割されずに、決定木学習法による選択器の生成は終了する。

【0066】そして、ステップS35において、閾値  $T_{HE}$  によって学習データ  $L$  を学習データ  $L_0, L_1$  に分割することは、図6に示す点A-B-C-Dによって囲まれる領域において、EBMT距離を示す横軸上の閾値  $T_{HE}$  の値に相当する点Kから縦方向に線100を引き、点A-B-C-Dによって囲まれる領域を、点C-D-J-Kによって囲まれる領域と、点A-B-K-Jによって囲まれる領域とに分割することに相当する。

【0067】また、ステップS35において、最大利得比  $MAX_{gain}$  を与えるのが  $E_{gain}$  であり、閾値  $T_{HE}$  によって学習データ  $L$  を学習データ  $L_0, L_1$  に分割することは、図6に示す点A-B-C-Dによって囲まれる領域において、TDMT距離を示す縦軸上の閾値  $T_{HE}$  の値に相当する点から横方向に線を引き、点A-B-C-Dによって囲まれる領域を2つの領域に分割することに相当する。

【0068】ステップS40, S50の各々において、学習データ  $L_0, L_1$  の各々がさらに2つの学習データ  $L_{01}, L_{02}; L_{11}, L_{12}$  に分割されると、上述した方法と同じ方法によって線101, 102を順次引く。そして、学習データ  $L_{01}, L_{02}$  および学習データ  $L_{11}, L_{12}$  のいずれかの学習データ  $L_{01}, L_{02}$  (または学習データ  $L_{11}, L_{12}$ )において、さらに、学習データの分割が行なわれると、線103, 104を順次引く。その結果、点H-M-K-Iによって囲まれる領域および点F-E-M-Gによって囲まれる領域が生成される。

【0069】この場合、分割された学習データにおいて、評価データのメンバーの値が全て同じ値になるか、演算された最大利得比  $MAX_{gain}$  が「0」以下になると、その分割された学習データの翻訳に適した翻訳システムの決定は終了する。最大利得比  $MAX_{gain}$  が「0」以下であるとき、対象とする学習データの翻訳に適した翻訳システムを決定する動作が終了することとしているのは、最大利得比  $MAX_{gain}$  が「0」以下になると、対象とする学習データを構成するTDMT距離  $T$  とEBMT距離  $E$  とによって決定される領域をさらに分割することができないからである。

【0070】このように、対象とする学習データを構成する評価データのメンバーの値が全て同じ値になるか、対象とする学習データを構成するTDMT距離  $T$  とEBMT距離  $E$  とによって決定される領域をさらに分割することができなくなるまで、学習データ  $L$  をツリー状に次々と分割することによって、TDMTによる翻訳に適した領域とEBMTによる翻訳に適した領域とに分割することができる。

【0071】図6に示す例では、点E-F-G-H-I-J-D-Cによって囲まれる領域は、EBMTによる翻訳に適した領域であり、点A-B-E-F-G-H-I-Jによって囲まれる領域は、TDMTによる翻訳に適した領域である。

【0072】従って、自動選択器60は、決定木学習法により図6に示すような各翻訳システムに適した領域に分割されたデータ（「選択器」と言う。）を受取ることによって入力データに適した翻訳システムを用いて入力データを翻訳できる。

【0073】決定木学習法による翻訳システムの選択を実際に行なった例について説明する。表1は、学習セッ

トおよび評価セットについて、各入力発話に対する TDMT および EBMT の出力を、評価者が一対比較した結果を示す。

【0074】

【表1】

	学習セット	評価セット
EBMT 勝	189	189
TDMT 勝	175	186
同等	144	135

【0075】表1に示すように、比較結果は、EBMT 優位（表1中の「EBMT勝」）、TDMT 優位（表1中の「TDMT勝」）および同等によって表わされる。

EBMT (DP Distance)	TDMT (Syntax Score)	Input	Teacher
		Result of Paired Comparison Method	
1	4.57		TDMT
0	0		TDMT
0.2	25		EBMT
0.2	0.67		TDMT
0	0.83		EBMT
:	:		:
0	1.33		EBMT

【0078】また、表3は、学習された決定木により自動選択を行なった結果を示す。

【0079】

【表3】

Decision Tree	EBMT	人間による選択	
		EBMT	TDMT
	EBMT	135(36%)	36(9.6%)
	TDMT	54(14.4%)	150(40%)

【0080】表3中の数字は、評価文の数を表わしており、括弧内の数字は評価セット全体（375文）に対する割合を表わしている。また、表3の下線部が正しく選択された結果を表わしており、正しく選択される割合は76%となっている。

【0081】図7は、表1～表3に示した場合を図示したものである。横軸はEBMT距離であり、縦軸はTDMT距離である。なお、直線105は、従来の方法によってTDMTによる翻訳に適した領域とEBMTによる翻訳に適した領域とに分割する場合の境界線である。また、「●」は人が入力データの翻訳に適している翻訳システムとしてTDMTを選択したことを表わし、「▲」は人が入力データの翻訳に適している翻訳システムとしてEBMTを選択したことを表わす。さらに、領域110～112がTDMTによる翻訳に適した領域であり、それ以外の領域113はEBMTによる翻訳に

また、表1中の数字は実際の文数を表わしている。元々の学習セットは508文、評価セットは510文からなるが、「同等」と判断されたデータについては、自動選択を行なう場合に、どちらが選択されても翻訳品質に影響がでないため学習および評価には用いていない。実際の学習および評価に用いたのは、一対比較により優劣が決定されたデータ（表1中の下線部）であり、学習セットは364文、評価セットは375文である。

【0076】表2は、決定木学習データの構造を示す。

10 EBMT距離、TDMT距離、および人間による評価結果が示されている。

【0077】

【表2】

適した領域である。

【0082】TDMTのプロットは、殆ど、領域110～112に含まれており、上述した決定木学習法は人が判断した場合とよい一致を示す。但し、EBMTのプロットも、領域112に含まれており、決定木学習法による翻訳システムの選択には、若干の誤りがあることがわかる。

【0083】全体として、決定木学習法による翻訳システムの選択が人間による評価結果と一致する割合は上述したように76%である。

【0084】従来は、TDMTによる翻訳に適した領域および/またはEBMTによる翻訳に適した領域への分割は直線105によってのみ可能であり、2つの領域の境界は直線的に決定されていたが、決定木学習法を用いることにより、2つの領域の境界は階段状に決定できるようになった。これにより、各種の入力データに適した翻訳システムを選択可能になった。

【0085】図8は、EBMT単体、TDMT単体、EBMTとTDMTとを決定木学習法により選択した場合、および人がEBMTとTDMTとを選択した場合の翻訳性能を、TOEICスコア685のTOEIC受験者により評価した結果を示す。システムによる翻訳がTOEIC受験者による翻訳よりも優れている場合（図8中の「MT勝」）、システムによる翻訳とTOEIC受験者による翻訳が同等の場合（図8中の「Even」）、およびTOEIC受験者による翻訳が翻訳シス

テムによる翻訳よりも優れている場合（図8中の「Human勝」）の3種類の結果が得られる。

【0086】図8の横軸において、左からEBMT単体での評価結果、TDMT単体での評価結果、決定木学習法により選択を行なった場合の評価結果、および人間が選択を行なった場合（すなわち、理想的な翻訳システムの選択を実現できた場合）の評価結果を示す。

【0087】図8においては、EBMT単体およびTDMT単体は、それぞれ、評価結果の割合がほぼ同じであり、ほぼ同等の翻訳性能である。決定木学習法により翻訳システムを選択した場合は、人間により翻訳システムを選択した場合よりも劣るが、EBMT単体またはTDMT単体の場合と比較して、翻訳システムによる翻訳が優位となる文の割合が増加し、TOEIC受験者による翻訳が優位となる文の割合が減少している。従って、決定木学習法により翻訳システムを選択し、EBMTとTDMTとを併用することにより、EBMTまたはTDMTを単体で用いた場合よりも翻訳性能が改善される。

【0088】上記においては、TDMTおよびEBMTという2つの翻訳システムから入力データの翻訳に適した翻訳システムを決定木学習法により選択する場合について説明したが、この発明においては、一般にm（mは自然数）個の翻訳システムから入力データの翻訳に適した翻訳システムを決定木学習法により選択してもよい。

【0089】図9は、m個の翻訳システム1～mから決定木学習法により選択された翻訳システムを用いて入力データを翻訳する場合の概念図である。自動選択器60は、決定木学習法により生成された選択器に基づいて、翻訳システム1～mから入力データの翻訳に適した翻訳システムを選択し、その選択した翻訳システムにより入力データを翻訳して出力データを出力する。

【0090】このように、入力データを翻訳する際の翻訳システムの数が増加することにより、入力データの翻訳に適した翻訳システムの組合せは増加し、入力データをより正確に翻訳できる。

【0091】この発明によるプログラムは、図10に示すパソコンコンピュータで実行される。図10は、パソコンコンピュータ90は、データバスBSと、CPU（Central Processing Unit）91と、RAM（Random Access Memory）92と、ROM（Read Only Memory）93と、シリアルインタフェース94と、端子95と、CD-ROMドライブ96と、ディスプレイ97と、キーボード98とを備える。

【0092】CPU91は、ROM93に格納されたプログラムをデータバスBSを介して読出す。また、CPU91は、シリアルインタフェース94、端子95およびインターネット網を介して取得したプログラム、またはCD（Compact Disk）99からCD-R

OMドライブ96を介して読出したプログラムをROM93に格納する。さらに、CPU91は、キーボード98から入力されたユーザからの指示を受付ける。

【0093】RAM92は、CPU91が各種の処理を行なう際のワークメモリである。ROM93は、プログラムおよびコード等を格納する。シリアルインタフェース94は、データバスBSと端子95との間でデータのやり取りを行なう。

【0094】端子95は、ケーブルによってパソコンコンピュータ90をモデム（図示せず）に接続するための端子である。CD-ROMドライブ96は、CD99に記録されたプログラムを読出す。ディスプレイ97は、各種の情報を視覚情報としてユーザに与える。キーボード98は、ユーザからの指示を入力する。

【0095】CPU91は、キーボード98から入力されたユーザのプログラム実行要求に応じて、データバスBSを介してROM93からプログラムを読出し、その読出したプログラムを実行する。このプログラムの実行により入力データはTDMTまたはEBMTによって翻訳され、その翻訳結果はディスプレイ97に表示される。これにより、人間はディスプレイ97を見て入力データの翻訳に適した翻訳システム（TDMTおよびEBMTのいずれか）を選択する。

【0096】また、CPU91は、RAM92を用いてEBMT距離およびTDMT距離を演算し、その演算結果をROM93に格納する。そして、CPU91は、キーボード98を介して評価データが入力されると、ROM93に格納されたTDMT距離およびEBMT距離をデータバスBSを介して読出し、上述した決定木学習法により翻訳システムの選択を行なう。翻訳システムの選択結果はディスプレイ97に表示される。

【0097】今回開示された実施の形態はすべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は、上記した実施の形態の説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

#### 【図面の簡単な説明】

【図1】 2つの翻訳システム（TDMTおよびEBMT）の概要を説明するための図である。

【図2】 複数の翻訳システムから入力データの翻訳に適した翻訳システムを自動選択する際の概念を示す概念図である。

【図3】 この発明によるプログラムがコンピュータに実行させるためのステップを説明するためのフローチャートである。

【図4】 図3に示すフローチャートのステップS30の詳細な動作を説明するためのフローチャートである。

【図5】 学習データの分割を説明するための図である。

【図6】 入力データの翻訳に適した翻訳システムの領域を示す図である。

【図7】 入力データの翻訳に適した翻訳システムの領域の例を示す図である。

【図8】 決定木学習法による翻訳システムの評価結果を示す図である。

【図9】 複数の翻訳システムを用いた翻訳の概念図である。

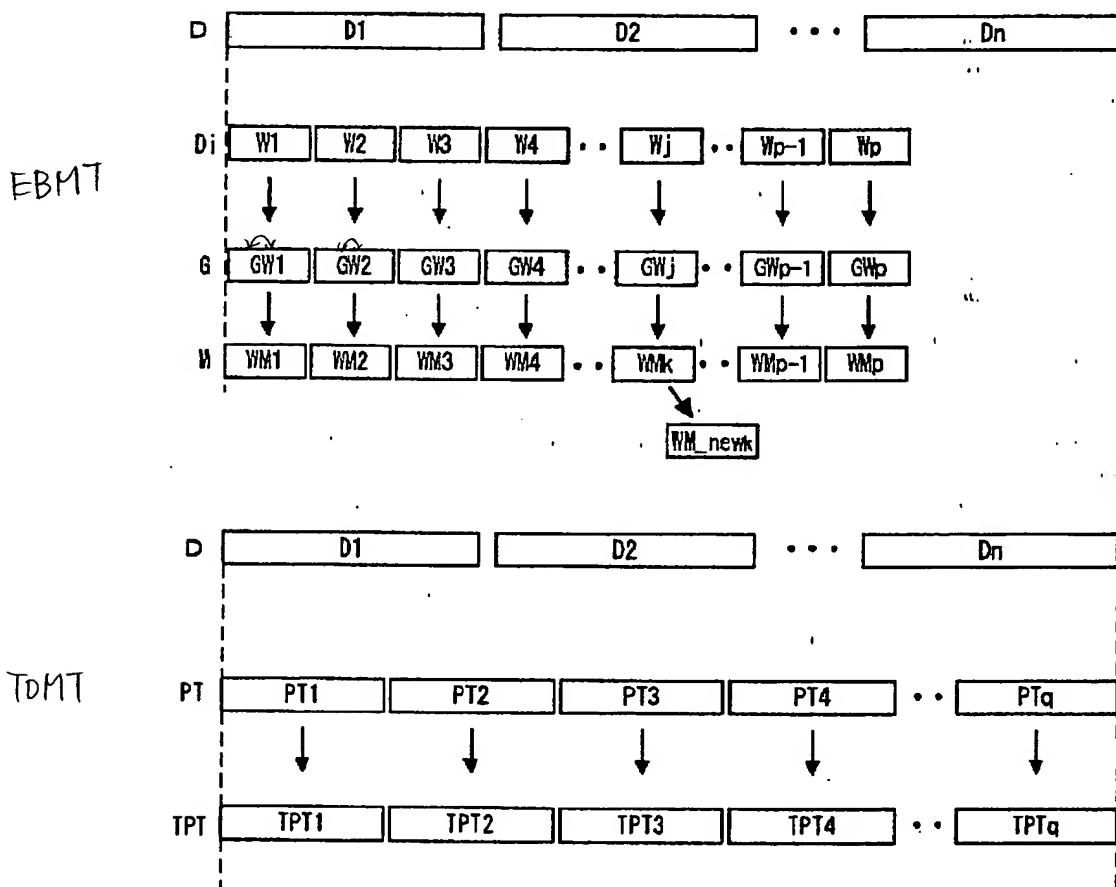
【図10】 パーソナルコンピュータの概略ブロック図

である。

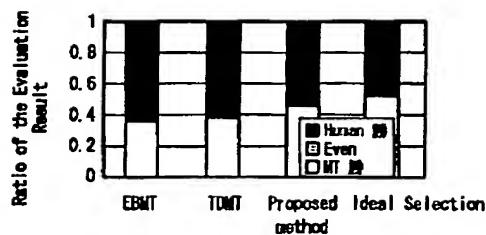
【符号の説明】

10 学習文、20 TDMT翻訳器、30 EBMT翻訳器、40 評価器、50 決定木学習器、60 自動選択器、90 パーソナルコンピュータ、91 CPU、92 RAM、93 ROM、94 シリアルインタフェース、95 端子、96 CD-ROMドライブ、97 ディスプレイ、98 キーボード、99 C D、100~106 線、110~112 領域。

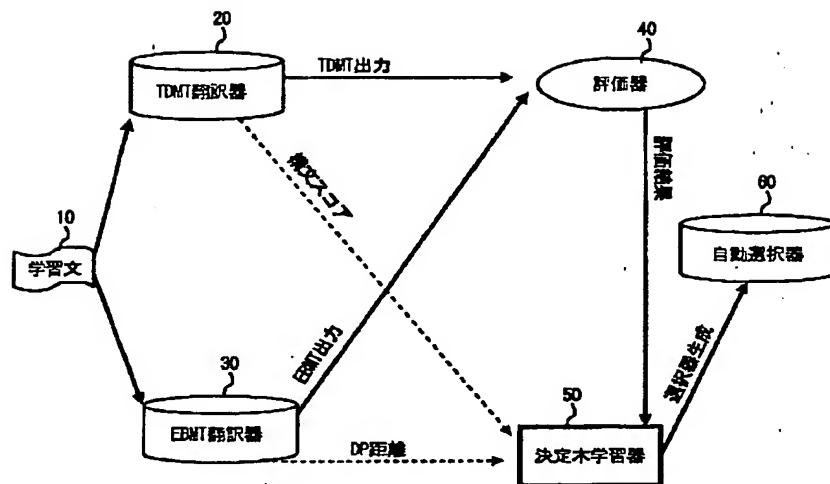
【図1】



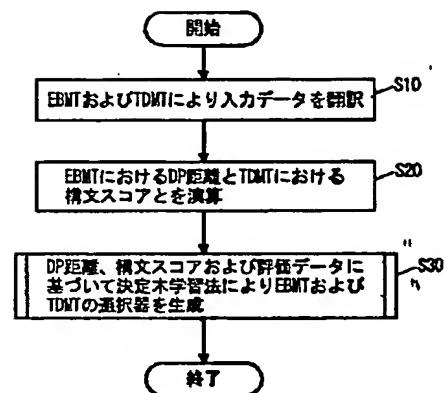
【図8】



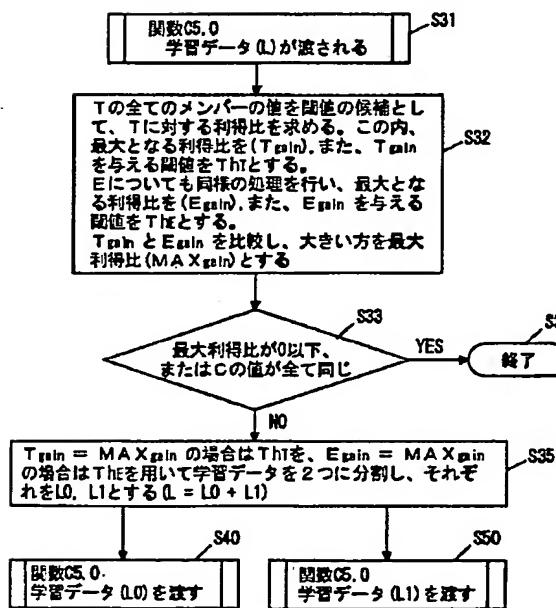
【図2】



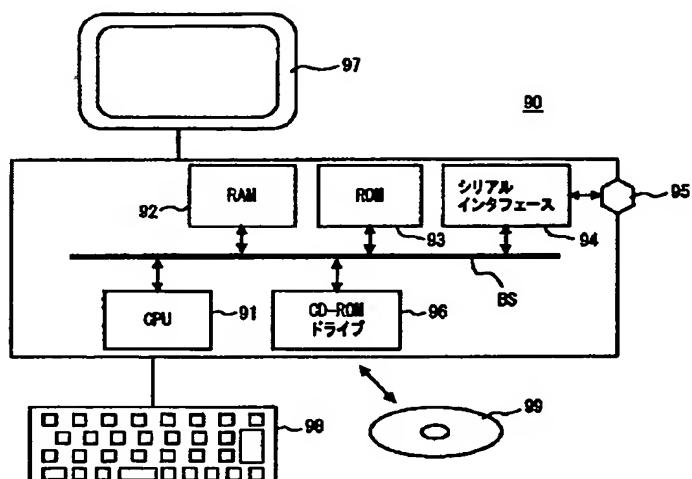
【図3】



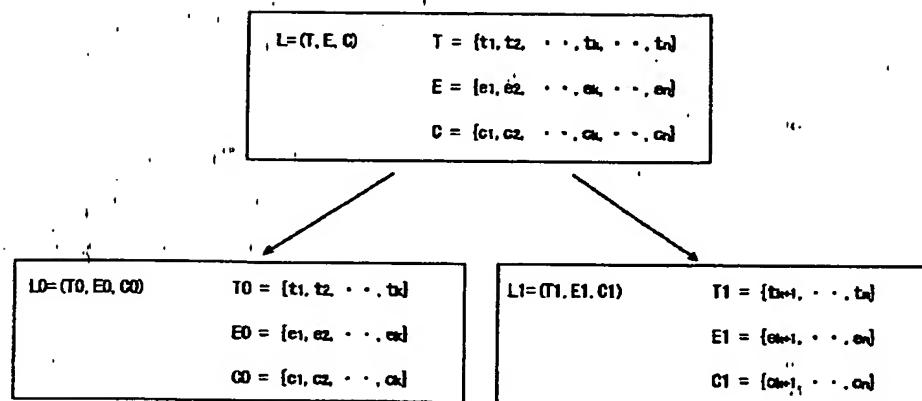
【図4】



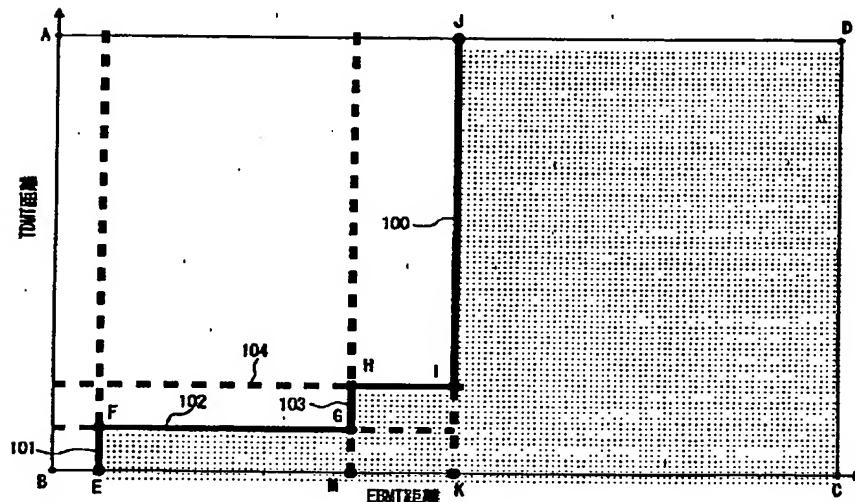
【図10】



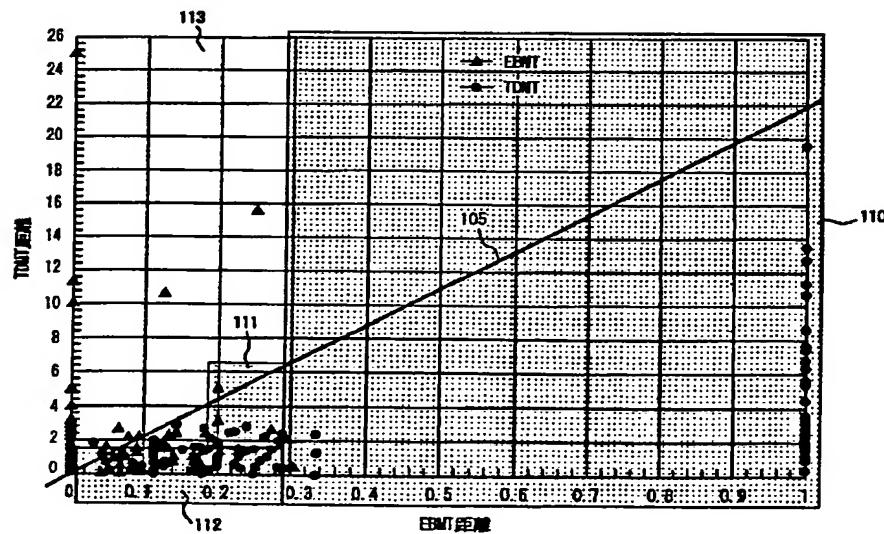
【図5】



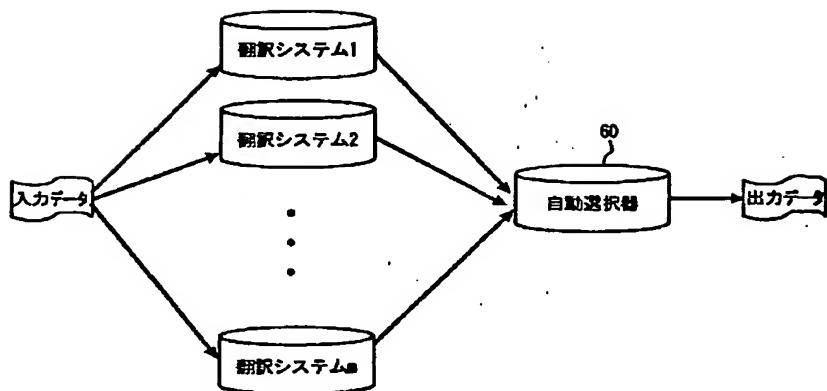
【図6】



【図7】



【図9】



フロントページの続き

(72)発明者 竹澤 寿幸

京都府相楽郡精華町光台二丁目2番地2  
株式会社国際電気通信基礎技術研究所内

(72)発明者 山本 誠一

京都府相楽郡精華町光台二丁目2番地2  
株式会社国際電気通信基礎技術研究所内  
F ターム(参考) 5B091 BA11 CD13 CD15

**This Page is Inserted by IFW Indexing and Scanning  
Operations and is not part of the Official Record**

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- BLACK BORDERS**
- IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- FADED TEXT OR DRAWING**
- BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- SKEWED/SLANTED IMAGES**
- COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- GRAY SCALE DOCUMENTS**
- LINES OR MARKS ON ORIGINAL DOCUMENT**
- REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- OTHER:** \_\_\_\_\_

**IMAGES ARE BEST AVAILABLE COPY.**

**As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.**